What is our effect size: Evaluating the Educational Influence of a Web-Based Intelligent Authoring Shell?

Slavomir Stankov
Faculty of Natural Sciences,
Mathematics and Education
University of Split
Teslina 12, Split
Croatia
stankov@pmfst.hr

Vlado Glavinić
Faculty of Electrical Engineering and
Computing
University of Zagreb
Unska 3, Zagreb
Croatia
vlado.glavinic@fer.hr

Ani Grubišić
Faculty of Natural Sciences,
Mathematics and Education
University of Split
Teslina 12, Split
Croatia
ani@pmfst.hr

***Abstract* – An educational system can be interpreted as a community where students and teachers are involved in the process of learning and teaching. Present-day educational systems present their users (teachers and students) an intelligent environment in order to enhance the learning and teaching process. Specifically, intelligent tutoring systems (ITSs) are computer systems designed for support and improvement of the learning and teaching process in a freely chosen knowledge domain. The goal of ITS developers is to build such systems that will create individualized instruction to get as close as possible to the 2-sigma boundary. As acquisition of knowledge is often an expensive and time-consuming process, it is important to know whether it actually improves student performance. In this paper we present some results on the evaluation of a Web-based ITS. Within this context we measure its educational effectiveness in augmenting students' accomplishments for a particular knowledge domain using the effect size as the metrics. By so doing we determine whether and in which degree an ITS increases students performance and can thus be an adequate alternative for human tutors.**

I. INTRODUCTION

Progresses in information and communication technology (ICT) as well as the latest developments in education technology have both enabled encouraging opportunities for introducing a new paradigm named e-learning. This paradigm ensures a learner centered, interactive, easy-to-access, flexible and distributable computer-supported environment [1]. However, the question that naturally arises when introducing such new artifacts is how effective and efficient they are with respect to traditional solutions. Hence an appropriate measuring apparatus based on a particular metric should be devised and applied.

In a widely quoted research Bloom [2] had compared student learning under three different forms of instruction: conventional learning, mastery learning and tutoring. Using the standard deviation of a control group (attending a conventional form of instruction), he found that the average tutored student was about two standard deviations (2-sigma) above the average control group one. In other words, tutoring improved the achievement of 50th percentile students to that of 98th percentile students. This research had subsequently started an avalanche of research seeking ways of accomplishing this result under more practical and realistic conditions than one-to-one tutoring with human teachers. One possible solution for, as stated by Bloom, the 2-sigma problem, is the usage of Intelligent Tutoring Systems (ITS), which provide each student with a learning experience similar to the ideal one-to-one tutoring. ITSs are computer systems that have both a clear representation of knowledge and diagnosis of students' errors, and can interact with students, guide the learning and teaching process to their needs and are less expensive than human tutors. The goal of ITS developers is to build systems which will create individualized instruction, accommodating student characteristics to more effectively promote student learning and to get as close as possible to the 2-sigma boundary.

On the other hand, according to Fletcher's results of a meta-analysis on the use of technology based instruction [3] it follows a steady progress toward Bloom's figure, which can thus be taken as the targeted effect size. In fact, as shown in Fig. 1, the effect size of 0.84 for ITSs indicates an improvement from the 50th to the 80th percentile achievement. The effect size of 1.05 for recent intelligent tutoring systems indicates an improvement from 50th percentile to 85th percentile educational achievement.

All instructional software should be evaluated before it is used in the educational process. We, like many other ITS developers, have become so involved in making our system work that we have forgotten our original goal: to build a tutoring system that is as good as or even better than highly successful human tutors. Moreover, we have paid little attention to the process of evaluation. Since the major goal of an instructional system is to teach, its evaluation's main test is to determine whether students learn effectively from it [4]. Although the ITS evaluation is a costly and time consuming involvement, it is necessary because it is the only way to find out what is the educational influence of an ITS on students.

In this paper, we present research on a specific web-based intelligent authoring shell, Distributed Tutor Expert

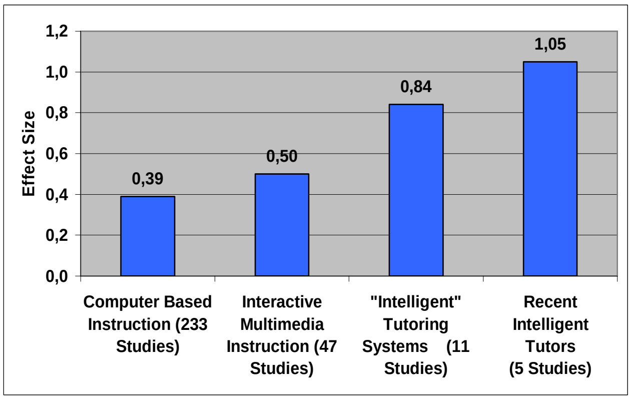


Fig. 1: Some Effect Sizes for Technology-Based Instruction [3]

System (DTEx-Sys) [5], which will provide an answer to the question “What is the educational influence of DTEx-Sys?” The objective of our research is to explore the effectiveness of DTEx-Sys, measured by students’ achievement in particular domain knowledge. Hence, the main task is to see if DTEx-Sys significantly increases the students performance in understanding certain domain knowledge, when comparing with traditional learning and teaching, and to see if it can be an adequate alternative to human tutors. We have to emphasize that DTEx-Sys so far has not been evaluated in this way, and we have no information about its effectiveness and its educational influence.

After a brief description of DTEx-Sys functionality, we give an overview of existing evaluation methods as well as the methodology we have used in our evaluation process. Finally, the evaluation results along with our discussion are presented.

II. BACKGROUND

The major problems when developing intelligent tutoring systems are their expensive and time consuming development process. In order to overcome those problems another approach has been chosen, namely to create particular ITSs from flexible shells acting as program generators. Such authoring shells should indicate design usability and flexibility to allow different representations of problem areas and to enforce an ease-of-use when developing an ITS for a particular problem area.

DTEx-Sys enables every student to learn at any moment and, what is even more important, to use system's services as long as she/he needs them for the purpose of achieving the required knowledge level [6], [7]. It is built upon the basic functionalities of the intelligent hypermedial authoring shell TEx-Sys (Tutor-Expert System) [8]. TEx-Sys supports teachers in the development of a series of intelligent tutoring systems for certain domain knowledge and enables students to learn, test themselves, consult the system for the obtained score, as well as receive from it advice for further work.

Within the TEx-Sys model, knowledge is represented by semantic networks with frames, whose basic elements are nodes and links. Nodes are used for presentation of domain knowledge objects, while links show relations between objects. Beside nodes and links, the system supports properties and frames (that are consisted of attributes and their respective values), along with property inheritance. Nodes can also have the following structure attributes: pictures, animations, slides, URL addresses and hypertextual descriptions.

The main functions defined by the TEx-Sys model are:

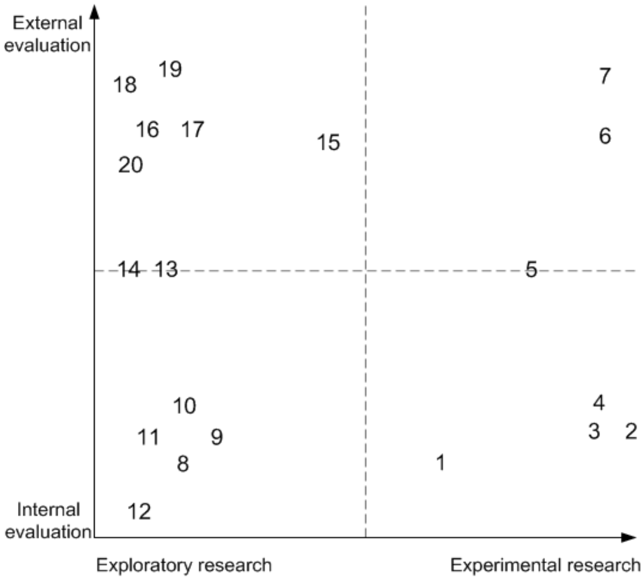
- authorization and registration to enable the legalization of work on the system;
- base knowledge creation of freely chosen domain (for teachers and in particular cases for students);
- learning and teaching upon developed domain knowledge base (for students);
- evaluation of a student's knowledge within a teaching scenario;
- access to achieved results (both for teachers and students);
- evaluation of a student's knowledge through quizzes [9]

Our present research is in developing a new system, xTEX-Sys (eXtended Tutor-Expert System) [10] based on TEx-Sys model, which will incorporate proposed educational standards from a field of e-learning (for example, Sharable Content Object Reference Model - www.adlnet.org). In order to reach the Blooms 2-sigma target, xTEx-Sys enhanced by some extended functions for courseware development and learning management, will be improved according to results gained through this evaluation.

III. EVALUATION METHODS

Given the variety of educational system evaluation methods, it is not as easy to decide which one is appropriate in a particular context [11]. Basically, there are two main types of evaluation methods [12]: formative and summative. *Formative* evaluation occurs during design and early development of a project. It is often a part of a software engineering methodology where it is used to obtain information needed for modifying and improving a system’s functionality. *Summative* evaluation is concerned with the evaluation of completed systems and tends to resolve, for e.g., such questions as: "What is the educational influence of an ITS on students?", "What does a particular ITS do?", "Does an ITS fulfill the purpose for which it was designed?", "Does an ITS result in predicted outcomes?"

All evaluation methods, irrespective of their type, are classified along two dimensions (Fig. 2.) [11]. The first dimension focuses on the degree of evaluation covered by the evaluating method. If the method only concentrates on testing a component of a system, it can be considered suitable for internal evaluation. If the method evaluates



1. Proof of Correctness 2. Additive experimental design 3. Diagnostic accuracy 4. Feedback/instruction quality 5. Sensitivity Analysis 6. Experimental research 7. Product evaluation 8. Expert knowledge 9. Level of agreement 10. Wizard of Oz experiment 11. Performance metrics 12. Internal evaluation 13. Criterion-based 14. Pilot testing 15. Certification 16. Outside assessment 17. Existence proofs 18. Observation & qualitative classification 19. Structured tasks & quantitative classification 20. Comparison studies

Fig. 2: Classification chart for evaluation methods [11]

whole system, it is suitable for external evaluation.

The second dimension differentiates between experimental research and exploratory research. Experimental research requires experiments that change the independent variable(s) while measuring the dependent variable(s) and require statistically significant groups. Exploratory research includes in-depth study of the system in a natural context using multiple sources of data, usually where the sample size is small and the area is poorly understood.

Experimental techniques are often used for summative research, where formal power is desired and where overall conclusions are desired. Common in psychology and education [4], experimental research is suited to educational systems, including ITSs, because it enables researchers to examine relationships between teaching interferences and students' teaching results, and to obtain quantitative measures of the significance of such relationships.

Different evaluation methods are suitable for different purposes and development of an evaluation is a complex process. In a variety of different experimental designs, we have decided to use control group experimental designs that enable determining the effects of particular factors or aspects of the evaluated system.

IV. EVALUATING THE EDUCATIONAL INFLUENCE OF DTEX-SYS

An evaluation answers the questions for which it was designed, hence the first step in research design is the identification of a research question. Hypotheses can be formed after identifying a research question, which must be testable, concerned with specific conditions and results, and possible to confirm or deny on the basis of those conditions and results. An evaluation methodology is then defined to enable the researcher to examine the hypothesis. When a practical, suitable evaluation method has been found to answer the research question, the researcher can carry out the study and analyze data gathered through the study. Ideally, if results do not confirm the research hypothesis, researchers should be able to suggest possible explanations for their results.

We decided to use *effect size* as a metric because it is commonly used in other evaluation studies, hence enabling us to compare our system with other evaluated systems [13]. An effect size is an index of the magnitude of a research result, such as the strength of the relationship between two variables or the amount of change produced by an intervention [14]. There are four types of effect size: standardized mean difference, correlation, explained variance, and interclass correlation coefficient, according to [13]. For determining group differences in experimental research, they recommended the use of standardized mean difference. The standardized mean difference [13] is calculated by dividing the difference between experimental and control group means by the standard deviation of the control group. The following formula is used for the calculation of this standardized score:

$$\Delta = \frac{\overline{X_e} - \overline{X_c}}{s_c}, \qquad (1)$$

where $\overline{X_e}$ = mean of the experimental group; $\overline{X_c}$ = mean of the control group; s_c = standard deviation of the control group. The mean or arithmetic average is the most widely used measure of central tendency, and the standard deviation is the most useful measure of variability, or spread of scores. Effect sizes can also be computed as the difference between the control and experimental posttest mean scores divided by the average standard deviation. According to [12] the effect size can be calculated using this formula:

$$\Delta = \Delta(post\text{-}test) - \Delta(pre\text{-}test). \qquad (2)$$

E.g., if the effect size for some imaginary computer-based treatment is 0.5 we would say that the experimental group outperformed the control group by 0.5 standard deviations. It can then be said that the typical student in an experimental group would perform at the 69th percentile (percentile is any of the 99 numbered points that divide an ordered set of scores into 100 parts each of which contains one-hundredth of the total) on the computer-based examination (see Table I. and Fig.3), while the typical student from the control group would perform at the 50th percentile (69% of the area under the standard normal

TABLE I

SOME AREAS UNDER NORMAL CURVE

Effect size	*Percentile*
0.0	50
0.1	54
0.2	58
0.3	62
0.4	66
0.5	69
0.6	73
0.7	76
0.8	79
0.9	81.6
1.0	84.1
1.1	86.4
1.2	88.5
1.3	90.3
1.4	91.9
1.5	93.3
1.6	94.5
1.7	95.6
1.8	96.4
1.9	97.1
2.0	97.7

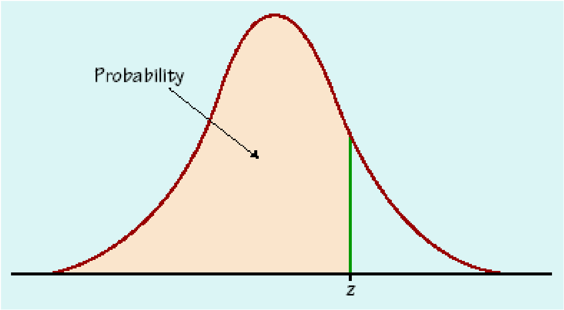


Fig. 3: Standard Normal Curve

curve falls below 0.5). The magnitude of the effect size is an indication of computer-based instruction effectiveness compared to traditional instruction.

A. Process of Evaluation

For purposes of DTEx-Sys evaluation, thirty three students taking the *Introduction to computer science* course were randomly and equally divided into *Control group* (11 students), *Tutoring group* (11 students) and *Experimental group* (11 students). The *Control group* was involved in traditional learning and teaching process; the *Experimental group* was asked to use system DTEx-Sys and the *Tutoring group* was tutored by human tutors (students from the *Tutoring group* were divided into four subgroups of 2-3 students and those subgroups were tutored by human tutors). All three different types of treatment were scheduled for two hours weekly throughout one semester (2hr/week x 15 weeks = 30 hours/semester).

All three groups underwent a 45-minute paper-and-pen pre-test that was distributed at the start of the course. Also, all three groups underwent a 60-minute paper-and-pen post-test that was applied two weeks after the end of the course. Their results were scored on a 0-100 scale. The pre-test enabled obtaining information on the existence of statistically significant differences among groups concerning student's foreknowledge. Post-test has enabled obtaining information on the existence of statistically significant difference among groups concerning evaluation influence of the DTEx-Sys. The *Experimental group* was asked to fill out a post-treatment anonymous questionnaire as a part of formative evaluation for the purpose of improving DTEx-Sys.

B. Analysis of results

The primary intention of this research is to evaluate the overall effectiveness and effect size of DTEx-Sys, so we computed and compared the *t-value* of means of gains of test scores among the three groups [14]. The *t-test* is the most commonly used method to evaluate the differences between two groups.

The *p-value* reported with a *t-test* represents the probability of error involved in accepting our research hypothesis about the existence of a difference. The critical region is the region of the probability distribution which rejects the null hypothesis. Its limit, called the critical value, is defined by the specified significance level. The most commonly used significance level is 0.05. The null hypothesis is rejected when either the *t-value* exceeds the critical value at the chosen significance level or the *p-value* is smaller than the chosen significance level. The null hypothesis is not rejected when either the *t-value* is less than the critical value at the chosen significance level or the *p-value* is greater than the chosen significance level. In our research, critical value for 11 degrees of freedom (number of students per group) and significance level 0.05 is 1.796. In the *t-test* analysis, comparisons of means and measures of variation in the two groups can be visualized in box-and-whisker plots. These graphs help in quickly evaluating and "intuitively visualizing" the strength of the relation between the grouping and the dependent variable.

We first checked whether groups' initial competencies were equivalent before comparing the gains of the groups. The mean pre-test score of the *Control group* was 50.91 with standard deviation of 15.10. The mean pre-test score of the *Experimental group* was 55.87 with a standard deviation of 25.26. The mean pre-test score of the *Tutoring group* was 50.97 with standard deviation of 22.00. We have computed the *t-values* of pre-test means (see Table II. and Fig. 4.) that enabled us to determine that there is no reliable difference between any two groups.

Then we have stated our hypotheses H1: "*There is no significant difference between the Tutoring and the Experimental group*" and H2: "*There is significant difference between the Control and the Experimental group*".

Next, the gain scores from pre-test to post-test were compared. The mean of the *Control group* was -5.18 with a standard deviation of 18.29. The mean of the *Experimental group* was 7.68 with a standard deviation of 13.48. The mean of the *Tutoring group* was 10.30 with a standard deviation of 17.50. We have computed *t-values* of means of gain scores (see Table III. and Fig. 5.) that enabled us to determine that there is a reliable difference between the *Control* and the *Experimental group* and between the *Control* and the *Tutoring group*. Also we have determined that there is no reliable difference among the *Experimental* and *Tutoring group*. The observed statistically significant difference implies that DTEx-Sys had a positive effect on the students' understanding of the domain knowledge. In other words, our hypotheses H1 and H2 are accepted.

The effect size is a standard way to compare the results of two pedagogical experiments. Hence, we calculate the effect size using (1)

TABLE II

T-TEST FOR PRE-TEST RESULTS

	t-test values	Significant difference

Control vs. Experimental	t=0.68
	p=0.26
		No

Control vs. Tutoring	t=0.01
	p=0.50
		No

Tutoring vs. Experimental	t=0.71
	p=0.25
		No

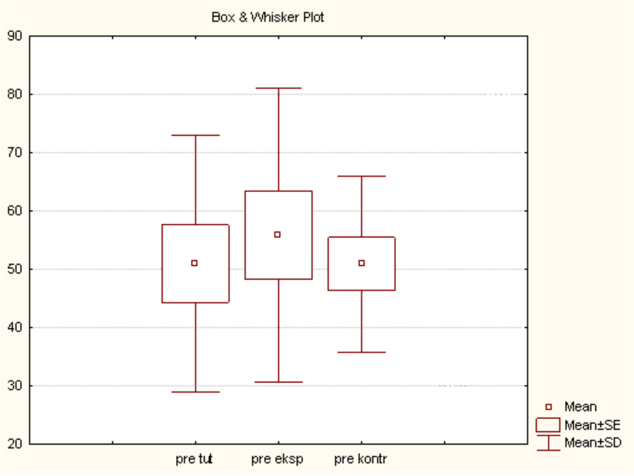


Fig. 4: Means for pre-test results

TABLE III

T-TEST FOR GAIN SCORES

	t-test values	Significant difference

Control vs. Experimental	t=2.41
	p=0.04
		Yes

Control vs. Tutoring	t=2.19
	p=0.05
		Yes

Tutoring vs. Experimental	t=0.53
	p=0.61
		No

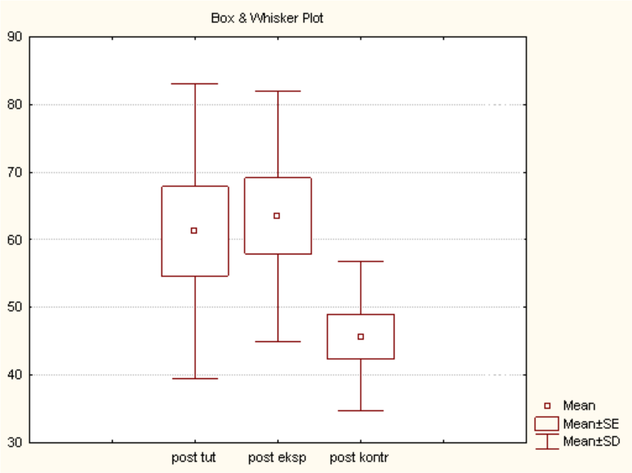


Fig. 5: Means for post-test results

$$\Delta = \frac{7.68-(-5.18)}{18.29} = 0.70, \qquad (3)$$

Therefore, we can say that the *Experimental group* outperformed the *Control group* by an amount of 0.7 standard deviations. The effect size of 0.7 indicates an improvement from 50th to 76th percentile achievement.

The average post-test score for the students in the *Experimental group* who used DTEx-Sys was 63.55, while the corresponding score for those in the *Control group* who did not was 45.73 (see Table IV.). The students who used the system scored, on the average, 17.82 points better on the post-test than those who did not. This difference is statistically significant (t=2.74, p=0.01). An analog comparison between the *Experimental* and the *Tutoring group* has shown that there is no statistically significant difference (t=0.26, p=0.79) between this two groups because the students in the *Experimental group* scored, on the average, only 2.28 points better on the post test than students from the *Tutoring group*.

TABLE IV

MEANS AND STANDARD DEVIATIONS

	pre-test	**post-test**	**gains**

Control	mean=50.91
	st.dev=15.10
		mean=45.73
		st.dev=11.04
			mean=-5.18
			st.dev=18.29

Tutoring	mean=50.97
	st.dev=22.00
		mean=61.27
		st.dev=21.85
			mean=10.30
			st.dev=17.50

Experimental	mean=55.87
	st.dev=25.26
		mean=63.55
		st.dev=18.50
			mean=7.68
			st.dev=13.48

The standard deviation for all three groups combined is 19.02. Hence, an increase of 17.82 points in the mean score is, using (2)

$$\Delta = 17.82/19.02 = 0.94 \qquad (4)$$

what is slightly less than the one-sigma increment in performance found for some other ITSs (cf. [3]).

Effect size can be calculated using different formulas and approaches, and its values can diverge. Using (3) and (4) we have calculated the effect size of DTEx-Sys and obtained different values. In our approach to evaluating the educational influence of a Web-based intelligent authoring shell, we have computed the average effect size in order to get a unique effect size that can be used in some meta-analysis studies

$$\Delta = (0.7 + 0.94)/2 = 0.82. \qquad (5)$$

This result, according to [3], positions DTEx-Sys in a category of standard intelligent tutoring systems. The effect size of 0.82 indicates an improvement from 50th to 79th percentile achievement that students from the *Experimental group* have shown when compared to those in the *Control group*.

The *Experimental group* was asked to fill out a post-treatment anonymous questionnaire for the purpose of software improvement. Results of the questionnaire have shown that the students, in majority, were comfortable with the interface which they considered to be easy to learn and comprehensible. When asked whether they would like to work more with the system and whether they would recommend the system to another student, the majority of them have answered “positively”.

VI. CONCLUSION

The evaluation described in this paper has enabled us to determine the size of educational influence of DTEx-Sys on students. So far, we have never conducted this kind of system evaluation and results we gained through evaluation process have encouraged us to continue our research in further developing web-based intelligent authoring shells.

The evaluation of the system indicates that the teaching strategy followed by DTEx-Sys is effective in accomplishing the task it was designed to perform. The *Experimental group* exceeded the *Control group* in every statistical test performed.

A significant difference between the *Control group* and the *Experimental group* has showed DTEx-Sys’s advantage over traditional learning and teaching. An insignificant difference between the *Experimental group* and the *Tutoring group* has showed DTEx-Sys’s competency in substituting human tutors.

DTEx-Sys effect size of 0.82 is slightly less than 0.84, a standard value for the intelligent tutoring systems (according to [3]). In spite of this success, we still have to improve the system as compared to human tutors, who on average yield a two standard deviation improvement over traditional classroom instruction.

VII. ACKNOWLEDGMENT

This work has been carried out within projects 0177110 *Computational and didactical aspects of intelligent authoring tools in education,* 0036033 *Semantic Web as Information Infrastructure Enabler* and TP-02/0177-01 *Web oriented intelligent hypermedial authoring shell,* all funded by the Ministry of Science and Technology of the Republic of Croatia.

VIII. REFERENCES

[1] Khan, B. H. "A framework for Web-based learning." in B. H. Khan (Ed.), *Web-based training.* Englewood Cliffs, NJ: Educational Technology Publications, 2001.

[2] B.S. Bloom „The 2 Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One Tutoring", *Educational Researcher*, 13, 1984, pp. 4-16.

[3] J.D. Fletcher „Evidence for Learning From Technology-Assisted Instruction", in H.F. O'Neal, R.S. Perez (Eds.), *Technology applications in education: a learning view*, Mahwah, NJ: Lawrence Erlbaum Associates, 2003, pp.79-99

[4] M.A. Mark and J.E. Greer „Evaluation methodologies for intelligent tutoring systems", *Journal of Artificial Intelligence and Education*, 4 (2/3), 1993, pp. 129-153.

[5] M. Rosić „Establishing of Distance Education Systems within the Information Infrastructure", M.Sc.Thesis, Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia, 2000 (in Croatian).

[6] M. Rosić, S. Stankov and V. Glavinić „DTEx-Sys – A Web Oriented Intelligent Tutoring System", in *Proceedings of Intelligent Conference On Trends in Communication - EUROCON 2001.*, Vol 2/2, 2001, pp. 255-258.

[7] M. Rosić, S. Stankov and V. Glavinić „Intelligent Tutoring Systems for Asynchronous Distance Education", in *Proceedings of Melecon'2000 10th Mediterranean Electrotechnical Conference*, Cyprus, Regional Communication and Information Technology, 2000, pp. 111-114.

[8] S. Stankov „Isomorphic Model of the System as the Basis of Teaching Control Principles in an Intelligent Tutoring System", PhD Diss, Faculty of Electrical Engineering, Mechanical Engineering and Naval Architecture, University of Split, Split, Croatia, 1997 (in Croatian)

[9] S. Stankov, M. Rosić, V. Glavinić "Using Quizzes in an Intelligent Tutoring System", in *Proceedings of International Summer School of Automation*, CEEPUS CZ_103, Maribor, Slovenia, 2001, pp. 87-91.

[10] S. Stankov, Principal Investigating Project TP-02/0177-01 *Web oriented intelligent hypermedial authoring shell,* Ministry of Science and Technology of the Republic of Croatia, 2003.

[11] A. Iqbal, R. Oppermann, A. Patel and Kinshuk „A Classification of Evaluation Methods for Intelligent Tutoring Systems", *Software Ergonomie '99 - Design von Informationswelten* (Eds. U. Arend, E. Eberleh & K. Pitschke), B. G. Teubner Stuttgart, Leipzig, 1999, pp. 169-181.

[12] D. Frye, D.C. Littman and E. Soloway „The next wave of problems in ITS: Confronting the "user issues" of interface design and system evaluation", in J. Psotka, L.D. Massey, S.A. Mutter & J.S. Brown (Eds.), *Intelligent tutoring systems: Lessons learned.* Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.

[13] N.Y. Mohammad „Meta-analysis of the effectiveness of computer-assisted instruction in technical education and training", doctoral dissertation, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, 1998.

[14] StatSoft, Inc. "Electronic Statistics Textbook", 2004, http://www.statsoft.com/textbook/stathome.html.

[15] P.L. Albacete and K.A. VanLehn, „Evaluating the Effectiveness of a Cognitive Tutor for Fundamental Physics Concepts" in *Proceedings of the 22nd Annual Meeting of the Cognitive Science Society*, 2000.

[16] J.R. Anderson, A.T. Corbett, K.R. Koedinger and R. Pelletier „Cognitive tutors: Lessons learned", *Journal of the Learning Sciences*, 4(2), 1995, pp. 167-207.

[17] Kinshuk, A. Patel and D. Russell „A multi-institutional evaluation of Intelligent Tutoring Tools in Numeric Disciplines", in *The Evaluation of Learning Technology Conference Proceedings (Ed. M. Oliver)*, University of North London, London, 2000, pp. 38-40.

[18] R.A. Wisher and T.M. Olson „The Effectiveness of Web-based Training", U.S. Army Research Institute for the Behavioral and Social Sciences, Research Report 1802, 2003.